The Royal Veterinary and Agricultural University
Food and Resource Economic Institute

Unit of Economics Working Papers 2001/14

Rational Inefficiencies.

Peter Bogetoft and
Jens Leth Hougaard

Rational Inefficiencies

Peter Bogetoft
Department of Economics
Royal Agricultural University, Denmark

Jens Leth Hougaard
Institute of Economics
University of Copenhagen, Denmark

September 2001
(Revised July 2002)

Abstract

In this paper, we suggest that inefficiency may be an indirect, on-the-job compensation to agents in an organization. We show how to use actual production data to reveal the trade-offs between different inefficiencies (slacks). Moreover, we discuss how to use this to improve productivity analysis as well as decision making and incentive provisions in organizations.

Keywords: Efficiency, Preferences, Incentives, Bargaining, Planning.

Correspondence: Peter Bogetoft, Department of Economics, Royal Agricultural University, Rolighedsvej 23, 1958 Frederiksberg C, Denmark. E-mail pb@kvl.dk

Acknowledgment: This paper was originally prepared for the Exclusive Workshop on DEA at University of Southern Denmark, Odense, 2001. It has subsequently been presented as a key-note address at both the First Hellenic Workshop on Efficiency, 2001 and the Asia Conference on Efficiency and Productivity Growth, 2002. It has also been presented at the second North American Productivity Workshop, 2002 and the Corporate Governance Workshop at rhus Business School, 2002.We which to appreciate the useful comments from numerous conference participants and from Timo Kuosmanen, Paul Wilson, Knox Lovell and Shawna Grosskopf in particular.

1 Introduction

Technical inefficiency is often interpreted as *waste* following the concept of X-efficiency by Leibenstien (1966, 1978). It means that too many inputs have been used to produce too few outputs. According to Leibenstein, X-efficiency is primarily caused by lack of motivation and lack of knowledge. If an inefficient firm does not motivate its employees sufficiently to save inputs and expand outputs, performance may be improved by redesigning the incentive structures. If inefficiency is caused by lack of information, performance may be increased by improving the markets for knowledge, learning etc.

Along similar lines, inefficiency may be related to sub-optimal decision procedures. According to work by Chris Argyris (cf. e.g. Leibenstein and Maital (1994)) the main source of technical inefficiency in firms is 'defensive behavior'. Employees are often reluctant to admit that their decisions were wrong – even though they themselves are aware of it. Thus the problem is not lack of information on how to optimize performance but rather lack of willingness to use this information in order to improve on procedures. According to the literature on organizational learning this problem lies in the structure of the organization.

On the other hand, technical inefficiency can also be interpreted as the result of *model mis-specifications*. Measured waste may simply reflect that not all inputs or outputs are accounted for, that heterogeneous inputs and outputs are pooled or that the assumed relationship between inputs and outputs is flawed. Taking this perspective, we should either refrain from making efficiency judgments or we should improve our modeling.

In this paper, we propose to explain or *rationalize* the presence of inefficiency. As such, we follow up on related thoughts by Stigler (1976) who argues against Leibenstein's concept of X-efficiency since "Leibenstein does not attempt to understand the allocation of "inefficient" resources, and hence does not see the necessity for attributing his X-inefficiency to specific inputs". In short, we try to do exactly that: we see inefficiency and in particular the allocation of inefficiency (slack) among different inputs as the result of a *rational choice* made by that firm.

There are many possible rationales for slack.

Measured inefficiency can be part of the *(fringe) compensation* paid to stake holders, e.g. the employees, the owners, the local community etc., and it may actually be the cheapest way to provide such compensation[1].

Inefficiency can also contribute to *incentives*. For example, the firm can pay its employees more than their opportunity costs in order to make them work efficiently out of fear of the harsh penalty associated with a dismissal for poor

[1]By the law of diminishing marginal rates of substitution, the marginal value of the first on-the-job excessive $-spending may exceed the value of the last off-the-job $-spending. Moreover, in a tax burdened society, the tax free on-the-job consumption has clear advantages over tax burdened private off-the-job consumption.

performance. It can create loyal employees and thereby reduce costly turnovers in the labor force[2]. Moreover, the cost of eliminating the inefficiency may exceed the worth of the waste. For example, extensive and costly monitoring may be necessary to eliminate slack or to substitute for a more high powered incentive scheme with harsh penalties. Similarly, painful arbitrations among groups and unions may be needed to eliminate inefficiency, and the arbitration costs may exceed the cost of the inefficiency.

In the organizational literature, technical inefficiency is also recognized as a possibly useful resource. Cyert and March (1963) define organizational slack as "the difference between total resources and total necessary payments". Thus, organizational slack can be interpreted as the existence of excess resources to produce a given output (technical inefficiency). In the literature there has been numerous discussions of the use of organizational slack as a *buffer* for uncertainty, as a means of de-coupling activities and hereby diminishing the information flow and coordination among subunits and as a necessity for providing resources for innovation, cf. e.g.. Galbraith(1974) and Stabler and Sydow(2001).

Technical inefficiency may also be caused by *rent seeking* behavior. In a regime where an organization is allocated a budget along with an output requirement, it may use resources in order to obtain better funding by manipulating the principal (or fonder). Part of the work force may, for example, be acting as lobbyists whose only contribution to the firm is to increase its probability of better funding (given the same output requirement). What appears to be technical inefficiency at a given point in time may hence be an attempt to maximize the expected profit in the long run[3].

The model and approach of this paper furthermore links up with the *bureaucracy* literature. Bureaucracies are popularly viewed as organizations which - absent a profit motive and free from the discipline of the market - pursue non-profit maximizing objectives. Niskanen(1971) argues that the managers derive utility from the size of the budget and that this may lead to production that exceeds the socially optimal size. The bureaucrat may well use a cost-minimizing mix of input but the level of resource use is socially inefficient. Migué and Bélanger(1974) combined this idea with the Williamson(1964) assumption that managers prefer larger staff. This means that the bureaucrat no longer uses a cost-minimizing mix of inputs given observed input prices. DeAllessi(1974) and others, on the

[2]Explanations along these lines are used also in the so-called efficiency wage theory - only in this case to explain why a firm may pay wages in excess of those that clear the market and why high wages and unemployment may coexist, cf. e.g. Weiss(1990).

[3]Originally rent seeking was used to explain that the welfare loss of monopoly was greater than assumed by the conventional theory of industrial organization. Here part of the loss in consumers surplus associated with a higher monopoly price is supposed to be transferred to the monopolist as rent (profit). However, if the government grants the right to obtain (and protect) a monopoly, firms will tend to invest resources in obtaining this right. Those resources are socially wasteful and should consequently be added to welfare loss connected with monopoly (cf. Tullock (1967)).

other hand, have argued that the managers will bias towards more capital intensive budgets. Lindsay (1976) suggests that they have a preference for more visible inputs (police cars rather than training for example) since they are easier to justify in the appropriation process. The competing hypothesis are formalized and tested on municipal data by Grosskopf and Hayes(1993). They find that the organizations are indeed somewhat inefficient but that there is no simple systematic in the way the inefficiency is introduced, i.e. via technical inefficiency or via allocative inefficiencies with preferences for staff or capital.

So the *general idea* of the rational choice perspective advocated here is that there are gains from inefficiency and therefore costs of improving efficiency. The gains are derived from the ability to offer on-the-job complementary payment, to improve incentives, and to ease planning, coordination and innovation in an uncertain environment. Also the gains may be associated with the fact that the units have other preferences, bureaucracy, or are exploiting special market conditions, rent seeking. Since we typically are ignorant about those gains and costs, we can not conclude that inefficiency as such should be eliminated. What we can do instead is to model the slack selection process and use this together with observed slack to make inference about the relative value of different types of slack. In this way, we try to *operationalize and measure the relative value of different types of slack* and we try to do so in a manner that can encompass several of the explanations above.

It is worth emphasizing that the apparently conflicting interpretations of inefficiency may all be simultaneously true. The idea that slack has value may obviously be interpreted as a mis-specification of the basic model in the sense that important outputs have been suppressed, cf. also below. The advantage of the rational choice perspective however is that it gives structure to the omitted variables and hereby discipline what can be excused as a model mis-specification. Similarly, the idea of slack having value does not necessarily mean that slack has value in a broader social context or that all slack has value. In a broader perspective, the slack may reflect a waste or misallocation due to inefficient social constructions (e.g. defects of taxation). Also, some inefficiency is probably "waste" in many contexts. Workers using sub-optimal work procedures may provide just as much effort and get just as little benefit from the work itself as worker using more efficient routines. Hence, the routines related inefficiency in this case is of no value[4].The distinction between valuable and non-valuable waste is a challenging one. Our rational choice perspective will put some restrictions on the distinction and some behavior may still be classified as wasteful as we will explain below. Still, in most cases it probably requires a deeper analysis of the nature of the slack to distinguish valuable form non-valuable slack. Put differently, while the traditional waste interpretation may be to harsh on organizations, we

[4]This is not to say that even this case cannot be rationalized. We may see the value of "waste" as a way more able workers are being rewarded for compared to less able ones.

may be too soft.

Observe that by introducing the perspective of inefficiency as a rational choice we challenge the *black-box view* of inefficiency underlying Leibenstein's tradition of X-efficiency which is, in fact, used in much of the modern literature on productivity analysis (including the well-known method of Data Envelopment Analysis (DEA), see e.g. Charnes, Cooper and Rhodes (1978)). As pointed out by Stigler (1976), the task of economists is not so much to point at inefficiencies but rather to explain why differences in efficiency occur. Once understood and modelled we are more qualified to deal with issues of efficiency. In the present paper we try to provide such a model which explains why we may observe differences in technical efficiency among otherwise similar production units. Stigler(1976) and Leibenstein(1966,78) both recognize the presence of differences in efficiency between firms, that is, the presence of X-inefficiency but whereas Stigler argues that the differences are the result of firms maximizing their individual value function, Leibenstein believes that the differences are caused by non-maximizing behavior. The danger of ascribing X-inefficiency to the presence of non-maximizing behavior is that the conclusion seems to be that society (or the shareholders) can gain by forcing the inefficient firm to be on the production frontier as in the underlying philosophy of DEA or productivity analysis in general. However, if what appears to be technical inefficiency at first sight actually is the result of rational and optimizing behavior of the firm such undertakings seem fruitless indeed, if not directly harmful over time.

The rational view suggests that the excess resource usages that show up as inefficiency are not just wasted. Rather, they are used to produce outputs that are not accounted for, e.g. a loyal pool of highly qualified and content employees, or to substitute for inputs that are not accounted for, e.g. a higher direct wage bill or higher turn-overs in the labor force. In this sense, inefficiency represents a measurement problem - what vanished in a crude model may play a useful role in a more detailed and comprehensive model. In this paper, we therefore accept that model mis-specifications may be at stake. We do, however, not forego the idea of measuring inefficiency on this ground. Rather, we suggest that the observed "inefficiencies" may proxy for the omitted aspects or values of slack and that this proxy may be used to predict more precisely the results of changes in the control instruments as well as to define other types of measurement, e.g. measures of allocative efficiency.

In terms of *methodology*, we will suggest a rather general model of the slack-selection which we can then calibrate on observed data by invoking the rationality hypothesis. We do so within a framework of *traditional production economics* as in e.g. Debreu (1951), Farrell (1957), Färe (1988), Koopmans(1951) and Shephard (1970). The paper is also related to much of the work we have done combining *incentive theory* and productivity analysis theory, cf. e.g.. Bogetoft(1994ab,95,97,2000), in the sense that we try to model more explicit an economic model explaining the determinants of inefficiency. Related work is also Haskel and Sanchis(1995,2000).

In a more restricted setting in terms of technologies, they use an employer-employee bargaining model to understand the determinants of the Farrell measures of technical and allocative inefficiency. Compared to these agency related contributions, however, we here develop a framework and a set of formulations that are more aligned with traditional, multiple dimensional production economics and less with the - from a production theoretical perspective - simple and stylized models used in the agency literature. Conceptually, the paper is also related to the "nonparametric production analysis" or "nonparametric tests of optimizing behavior" literature as Varian (1984, 1985, 1990) calls it. This approach, rooted in the theory of revealed preference of Samuelson(1947) and the work of Afriat (1972) and Hanoch and Rotschild (1972), has many technical similarities to DEA as pointed e.g. out by Banker and Maindiratta (1988) and Färe and Grosskopf (1995), and is not void of applications, cf. e.g.. Chavas and Cox(1990, 1992, 1995). A fundamental difference to much of modern productivity analysis literature and DEA in particular is that where DEA is mostly concerned of identifying and measuring the degree of "inefficiency/irrationality", this literature is mostly occupied with making inference of producer behavior in the spirit of the revealed preference approach after the empirical test first confirms rationality (consistency with a given optimization hypothesis, i.e. full efficiency of all DMUs), see Varian(1984). If departures from rationality are observed, they are typically "explained away" by measurement errors (Varian, 1985) or omitted variables (Varian, 1988).

To formalize and operationalize the rational perspective, we introduce in Section 2 a model of the values derived from different types of remunerations, off-the-job as well as on-the-job. In Section 3 we define some key aspects of the corporate governance structure in terms of alternative planning or control regimes, one being price based and the other being restriction based. It is within these control systems that rational behavior leads to inefficiencies and we therefore need to understand these systems to make inference about the revealed values. In Section 4 and 5, we show how a rational Decision Making Unit (DMU) will respond in these systems. We use these responses together with observed behavior to estimate the value trade-offs in the DMU in Section 6. Some applications of the estimated value models in decision making, incentive provision and productivity analysis are discussed in Section 7 and some final remarks are provided in Section 8.

2 The Model

We consider a setting where a DMU has used inputs $x \in \mathbf{R}^p_+$ to produce outputs $y \in \mathbf{R}^q_+$.

To simplify the exposition, we focus on the input space.[5] A technology is

[5]Generalizations to cases were both excess inputs and output shortage may have value is

therefore defined by the *input requirement set*

$$L = \{x \in \mathbf{R}^p_+ | x \text{ can produce } y\}$$

Let L be non-empty, closed, free disposable (comprehensive, i.e. $L + \mathbf{R}^p_+ \subseteq L$), and for ease of exposition convex[6].

We define the *efficient* subset of L as $F(L) = \{x' \in L \mid \forall x" \in \mathbf{R}^p : [x" \leq x', x" \neq x'] \Rightarrow x" \notin L\}$. The efficient subset represents the technically efficient production plans. It involves no "waste" or excessive consumption of inputs. Consider now the case where the DMU has used an *inefficient* input combination $x \in L \backslash F(L)$. We will think of this as a result of the DMU (or some agents inside the DMU) having value for on-the-job consumption. The DMU may have used the procedures and techniques associated with any production plan z that weakly dominates x, i.e. any *underlying production plan* $z \in L$ with $z \leq x$. The difference $x - z$ represents excess usage of inputs, or what we shall often refer to as consumption of *slack*

$$s = x - z \in \mathbf{R}^p_+.$$

With no further information, we cannot know exactly which plan z the DMU has used as the "underlying" production plan. We will introduce additional assumptions below and use these to make more specific predictions about z. Along the same lines, we cannot say exactly which slack vector the DMU has consumed, but we know, having observed x, that it belongs to the *slack possibility set* $S(x)$ *given* x defined by:

$$S(x) = \{s | s = (x - z), z \in L, z \leq x\}$$

as illustrated in Figure 1 below.

---

possible. On the output side, one can for example think of a baker eating some of his own donuts - or more relevant perhaps, the consumption of own products at farms.

[6]Much of what we do can be generalized to non-convex technologies. Convexity is however sed in the proofs of Proposition 2 and the last two parts of Corollary 2.

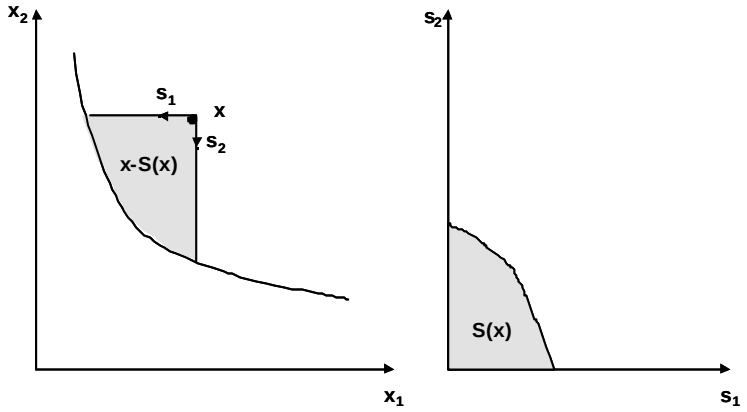

Figure 1: The Slack Possibility Set Given Actual Inputs

In addition to slack, the DMU is compensated in terms of direct, off-the-job consumption. We shall aggregate the off-the-job possibilities in a single index, referred to as profit π for simplicity.

We assume that the *DMU's value* depends on both off-the-job profit $\pi \in \mathbf{R}_+$ and on-the-job slack $s \in \mathbf{R}^p_+$. For some of the developments below, we simply need any ordinal preference relation with more preferred to less. For simplicity, however, we assume the existence of some underlying value function

$$U = U(\pi, s)$$

that is (strictly) increasing in both arguments, i.e. $(\pi', s') \geq (\pi, s)$ and $(\pi', s') \neq (\pi, s) \Rightarrow U(\pi', s') > U(\pi, s)$.[7]

When we postulate that the DMU has a value function it is of course an open question whether this value function is related to the shareholders or the employees (or even managers or workers). However, using a value function as proposed above this question becomes irrelevant: Assume that we model the shareholders utility. Clearly, (off-the-job) profit is important to the shareholders, but it is also in the interest of the shareholders to include concern for the employees (modelled by the slack component) as the company benefits from having a content and well motivated staff. As an example, many firms use slack (fringe benefits) in the competition for qualified labor. On the other hand, assume that we model the

[7]The analysis of this paper can easily be modified to work with weakly increasing value functions. The results in this case are similar except that there may be multiple solutions to our programs and that the properties we developed therefore hold for one as opposed to all possible solutions.

employees utility. Clearly, slack is important to the employees, but it is also in the interest of the employees to include profit in the utility function as they want the ongoing accept of the shareholders – and without profit firms close. Thus we may find that both profit and slack components are important no matter from which point of view we model the firm's "utility" and the trade-off between the two is left for the specific functional form.[8]

The DMUs ability to control off-the-job profit and on-the-job slack depends on the context. Whatever context, however, we assume that the DMU exploits its possibilities in a *rational* manner. If the DMU has the ability to choose the profit-slack vector from some subset Γ of $\mathbf{R}_+^{1+p}$, we assume that it does so to solve

$$\max_{(\pi,s)\in\Gamma} U(\pi, s).$$

Below, we focus on two contexts or *planning regimes*, one that uses price signals and another that uses quantitative restrictions. In both regimes, it is assumed that the DMU is requested to produce at least the output profile y with associated input requirement set L.

For analytical purposes, it is particularly useful to assume *preferential independence* between π and s. This means that the level of profit does not affect the trade-off between the different slack types and in particular that we can rewrite the value function as

$$U(\pi, s) = V(\pi, g(s))$$

where $g(.) : \mathbf{R}_+^p \to \mathbf{R}$ is an increasing function that aggregates the value of the different types of slack. This simplifies the analysis below because it allows us to separate the choice between on- and of-the-job consumption from the choice of the composition of on-the-job consumption. A similar separation of income and effort effects is commonly introduced in the agency literature.

Example 1: A straightforward possibility is to use an additive model like

$$U(\pi, s) = \alpha\pi + (1-\alpha)g(s), \quad \alpha \in [0, 1].$$

If $\alpha = 1$, we obtain a classical profit center with no internal incentive problems. If $\alpha = 0$ we obtain a classical employee / agent organization with no cost saving incentives.

[8] Leibenstein (1978) points at empirical evidence stating that monopoly leads to higher costs than competition - a 'loss' that is ascribed to X-inefficiency. Using our model, 'losses' from lack of competition can be rationalized in the sense that on-the-job slack will play a more dominant role maximizing the utility of a monopoly: Since the monopoly earns over-normal (off-the-job) profits shareholders are in general content leaving much more room for on-the-job slack to employees. A somewhat similar line of argument can be found in Parish and Ng (1972) who includes leisure in the utility function of the monopolist. As such a monopoly may very well be allocatively efficient in a broader sense (to be made precise in the following) and yet, on the surface, appear to be technically inefficient (or X-inefficient).

If, in addition, we assume that the function $g(\cdot)$ is also a simple additive function we get the objective function:

$$U(\pi, s) = \alpha_0 \pi + \sum_{i=1}^{p} \alpha_i s_i.$$

Note that this objective is similar to the one used in Bogetoft(1997,2000) in single inputs (cost) models. For multiple input cases, it is not entirely convenient because the linear structure typically leads to extreme, non-inner solutions to the problem of selecting appropriate slack combinations. □

However, it is easy to imagine more complex slack aggregation functions. We shall introduce some alternative value of slack aggregation functions in Section 6.

3 Planning Regimes

In this paper, we focus on two contexts or *planning regimes*. In both regimes, the DMU is requested to produce at least the output profile y with associated input requirement set L.

In the *price based regime*, the DMU is being allocated a certain *budget* or total payment $b \in \mathbf{R}_+$. Moreover, it is given *input prices* w at which inputs can be acquired. In this case, using inputs x leaves the DMU with *profit* $\pi = b - w \cdot x$. The DMU's aim is therefore to solve the following program

$$\begin{array}{ll} \max & U(\pi, s) \\ \pi, s, x & \\ \text{s.t.} & \pi \leq b - wx \\ & x - s \in L \\ & s \geq 0, x \geq 0 \end{array}$$

where x is the actual vector of inputs consumed. This program depicts the DMU under price control as choosing an input combination x and a slack vector s that are compatible with the output requirement $x - s \in L$. The DMU also picks the profit level which (given the use of inputs x) can be at most $b - wx$. Observe that we do not require the resulting off-the-job consumption π to be non-negative nor to exceed a certain minimum. In many cases, such "individual rationality" or "limited liability" constraints are of course relevant. They may however be encompassed in the present framework by assuming that the $U(\pi, s)$ value becomes prohibitively low for values of π not fulfilling such extra requirements.

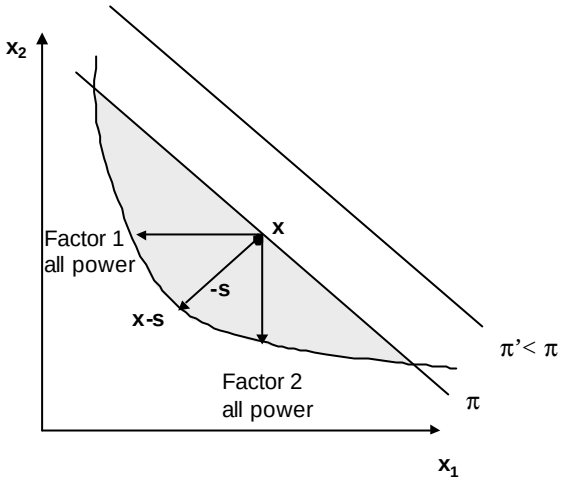

Figure 2: Price Based Planning

The DMU's decision problem under price based planning is illustrated in Figure 2. By decreasing the profit level $\pi = b - wx$, i.e. by moving the profit line towards north-east, the DMU's slack possibilities (as reflected by the shaded area) increases. For any given profit level, the choice of x and s reflects the relative weights associated with the different types of slack. If the first production factor, say doctors in a hospital, holds all the bargaining power relative to the other groups of health care personnel, we would expect all gains (slacks) to go to this group. This means that the DMU will choose $s = (s_1, s_2)$ proportional to $(1, 0)$ and as large as possible for given x. (We shall develop more precise predictions of the DMU's choices in the subsequent sections.)

In the *restriction based regime*, the DMU is given the right to use any input vector from a set of acceptable input combinations A. We assume that A is compact and that it does allow the production of y, that is $A \cap L \neq \emptyset$. Moreover, the DMU is given a direct payment or budget b with the interpretation that this amount can be consumed off-the-job. The DMU's aim is therefore to solve the following program

$$\begin{array}{ll} \max & U(b, s) \\ s, x & \\ \text{s.t.} & x \in A \\ & x - s \in L \\ & s \geq 0, x \geq 0 \end{array}$$

This program depicts the DMU under quantitative restrictions as choosing an acceptable production plan $x \in A$ as well as a slack vector s such that the

resulting resources $x-s$ can produce the required amount of output, i.e. $x-s \in L$.

The DMU's decision problem under restriction based planning is illustrated in Figure 3. In this case, the slack possibilities (as reflected again by all vectors commencing and ending in the shaded area) does not depend on the profit but only on the technology L and the planning restriction A. Off-the-job consumption (profit) is fixed and the only concern of the DMU will be to maximize slack taking into account - as before - the relative value of the different types of slack.

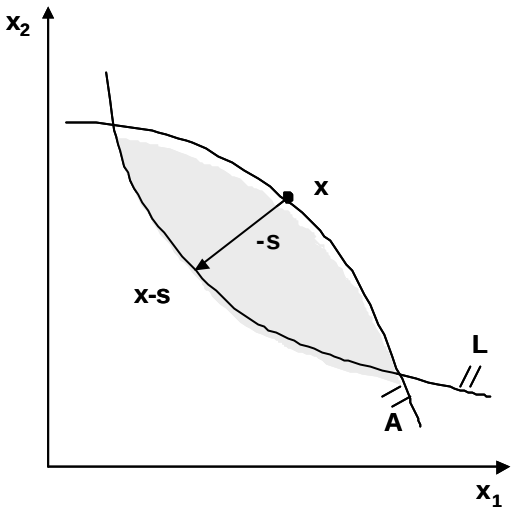

Figure 3: Restriction Based Planning

4 Production in a Price Based Regime

Consider first the price based regime, i.e. assume that the DMU is coordinated by output requirement y, budget b, and input prices w. The rational DMU will then choose profit π, total inputs x and slacks s that solve

$$\begin{array}{ll} \max & U(\pi, s) \\ \pi, s, x & \\ \text{s.t.} & \pi \leq b - wx \\ & x - s \in L \\ & s \geq 0, x \geq 0 \end{array}$$

An optimal underlying production plan $z = x - s$, is then an *allocatively efficient input combination* in the sense that it belongs to

$$X^a = \arg\min_{x \in L} w \cdot x$$

We record this as a Proposition.

Proposition 1 *In an optimal solution to the price based problem, the DMU will choose an allocatively efficient input combination as the underlying production plan* $z = x - s \in X^a$.

Proof

By changing the choice variables from π, s, x to s, z and by using that $U(\pi, s)$ is increasing in π, the DMU's program can be rewritten as

$$\begin{array}{ll} \max & U(b - w \cdot z - w \cdot s, s) \\ z, s & \\ \text{st} & z \in L, s \geq 0 \end{array}$$

Again by $U(.,.)$ increasing in the first variable (profit), the DMU will choose $z \in L$ to minimize $w \cdot z$, i.e. it will choose an allocatively efficient input combination $z \in X^a$. □

The content of Proposition 1 is illustrated in Figure 4.

The intuition and interpretation of Proposition 1 runs as follows: The DMU generates the necessary inputs in the cheapest possible way using $z = x^a \in X^a$. Whatever budget is left, $b - w \cdot x^a$, is then used for profit and slack. To emphasize this, note that given an optimal choice of the underlying production plan, the DMU's remaining decision can be formulated as the following *slack selection problem*

$$\max_{s \in \mathbf{R}^p_+} U(b - w \cdot x^a - w \cdot s, s)$$

In Section 6, we shall show how to use the observed solution to this problem to make at least partial inference about the DMU's trade-offs between profits and slack in the different input dimensions.

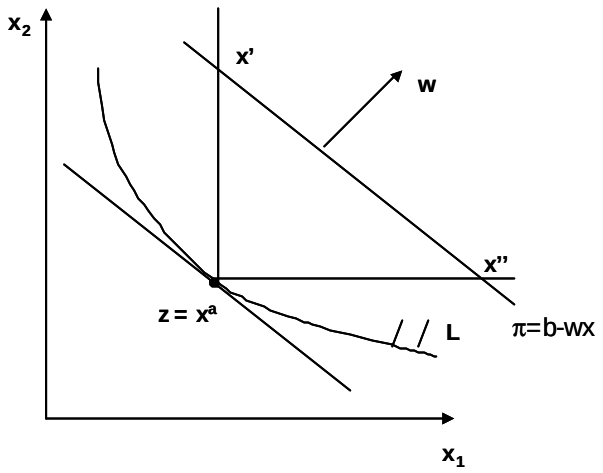

Figure 4: Choice of Underlying Production in Price Based Regime

From the point of view of productivity analysis, Proposition 1 shows that the traditional definition of allocative efficiency is useful even in the extended setting we consider here. Note that in our setting, prices reflect the market or control conditions of the planning regime and they do not necessarily reflect the values of the DMU. The values of the DMU are determined by the profit that it generates (which of course depends at least partly on w) and the slack that it consumes. The underlying preferences for slacks, say in labor vs capital, does not affect the underlying production plan $z = x^a$ - but it will affect the actual production plan x[9].

Although the choice of actual production plan x depends on the specific preferences for slack and profit, we can put simple constraints on the possible production plans a rational DMU can choose. From Proposition 1 we have that $z = x - s \in X^a$ and since $s \geq 0$, we get $x \in X^a + \mathbf{R}^p_+$. We record this a corollary.

Corollary 1 *In an optimal solution to the price based problem, the DMU will choose the actual production plan as* $x \in X^a + \mathbf{R}^p_+$.

That is, in order to obtain the 'optimal' slack possibility set the DMU must locate itself in a position where some point in X^a will dominate it; the exact location will of course depend on the specific value function of the DMU. This again gives rise to a new notion of allocative efficiency because we may, given our

[9]As indicated in the introduction, Grosskopf and Hayes(1993) test for differences between the input mix in x^a and x in a bureaucracy setting.

model above, infer that all DMU's in $X^a + \mathbf{R}^p_+$ are potentially allocatively efficient. On the other hand, DMU's that are not in this set are termed allocatively *in*efficient as they could have improved their slack possibilities (and hence their value from slack) without affecting profit, see also Section 7.3.

In other words, a rational DMU will take the following steps to decide how to produce the required amounts of outputs. First, given the prices w and the technology L, the DMU identifies the cost minimizing point(s) X^a that are able to produced the requested outputs. Actual input usage must therefore belong to $X^a + \mathbf{R}^p_+$, i.e. it must use at least the same as an allocatively efficient plan. Next, profit decreases as the iso-cost hyperplane moves upwards, and the trade-off between on- and off-the-job consumption determines how much profit should be forgone to allow for slack possibilities. Finally, given some profit level, maximization of the slack aggregation function determines the exact location on the segment of the iso-cost hyperplane that intersects $X^a + \mathbf{R}^p_+$, e.g. the segment x' to x'' in Figure 4.

5 Production in a Restriction Based Regime

Consider next the restriction based regime, i.e. assume that the DMU is coordinated by output requirement y, a off-job-consumption level b, and an acceptable set of inputs A. The rational DMU will then choose inputs x and slacks s that solve

$$\begin{array}{ll} \max & U(b,s) \\ s,x & \\ \text{s.t.} & x \in A \\ & x - s \in L \\ & s \geq 0, x \geq 0 \end{array}$$

The objective of the DMU is the same as in the previous regime. It seeks to maximize the value from off-the-job payment (profit) and on-the-job compensation (slack). Since the profit in this regime, where there are no possibilities to trade, is equal to the budget, we have used $U(\pi, s) = U(b, s)$ directly in the formulation of the DMU's problem. Note also that the acceptance set A is determined by the shareholders or management and that we therefore in the restriction based regime has a pure on-the-job slack selection problem.

Before introducing more specific regularities on A it is useful to observe that to maximize consumption of slack, the DMU will always choose an underlying production plan $z = x - s$ on the efficient part of L since otherwise it could increase at least some slacks without reducing others by simply moving z to one of the efficient points that dominates the original proposal. The same reasoning suggests that the DMU will always choose the actual inputs on the north-east part of A. To formalize this, we may for any $X \subseteq \mathbf{R}^p$ let $D(X) = \{x \in X | \forall\ x' \in$

$\mathbf{R}^p: \ [x' \geq x, \ x' \not= x] \Rightarrow \ x' \notin X\}$, i.e. $D(X)$ is the set of input vectors that cannot be increased in any dimension in the set X.

Observation: Any optimal solution in the restriction based regime satisfy $z = x - s \in F(L)$ and $x \in D(A)$

Proof: Assume that $z \notin F(L)$. Then by definition there exist an alternative z' such that $z' \leq z$ and $z' \neq z$. This means that $s' = x - z' \geq s = x - z$ and $s \neq s'$ and by $U(b, s)$ increasing in slacks s, we have a contradiction since $U(b, s') \geq U(b, s)$ such that (b, s) cannot be an optimal solution to the DMU's problem. The case $x \notin D(A)$ similarly leads to a contradiction. □

To further characterize the optimal choice of underlying production plan z, it is useful to introduce convexity. Assuming that A is convex, we can show that the optimal z - as in the price directed regime - is allocatively efficient, only now with respect to a set of prices w defined endogenously from A and the DMU's choice of actual production plan x. That is, the prices, which are absent as a control instrument in the restriction based regime, can be developed endogenously and they can be used to characterize the solution in much the same manner as in the price directed regime. We record this as a proposition.

Proposition 2 *In an optimal solution to the restriction based problem with A convex, there exists a set of prices $w \in \mathbf{R}^p_+$ such that x maximizes wx' over A and z minimizes wz' over $\dot{L}$*

Proof:

The DMU's choice problem in the restriction based regime can be rewritten as

$$\begin{array}{ll} \max & U(b, s) \\ s & \\ \text{s.t.} & s \in (A - L) \cap \mathbf{R}^p_+ \end{array}$$

By $U(b, s)$ being increasing in s, an optimal s will belong to the efficient part of $(A - L) \cap \mathbf{R}^p_+$. Moreover, by A and L convex, so is $(A - L) \cap \mathbf{R}^p_+$. It follows therefore by the weak separation result for convex sets that there exists a hyperplane with normal vector or associated prices $w \in \mathbf{R}^p_+$ such that the optimal s is separated form $(A - L) \cap \mathbf{R}^p_+$, i.e. such that

$$ws \geq ws' \qquad \forall s' \in (A - L) \cap \mathbf{R}^p_+$$

Let $x \in A$ and $z \in L$ be the optimal values of the actual and underlying inputs associated with s. Similarly, let $x' \in A$ and $z' \in L$ be the optimal values of the actual and underlying inputs associated with s'. In particular then $s = x - z$ and $s' = x' - z'$. It follows from the separation property that

$$w(x - z) \geq w(x' - z') \qquad \forall (x' - z') \in (A - L) \cap \mathbf{R}^p_+$$

By $w \geq 0$, it follows that

$$w(x - z) \geq w(x' - z') \qquad \forall (x' - z') \in (A - L)$$

and by considering the inequalities resulting when $z = z'$ and $x' = x$, respectively, we get

$$wx \geq wx' \qquad \forall x' \in A$$
$$wz \leq wz' \qquad \forall z' \in L$$

as desired. □

The content of Proposition 2 is illustrated in Figure 5 below.

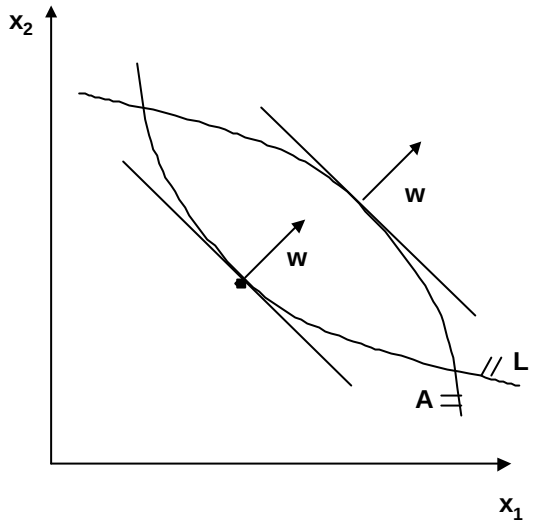

Figure 5: Choice of Underlying Production in Restriction Based Regime

We note that Propositions 1 and 2 are closely related. The DMU's problem in the price based regime can be decomposed into a problem of determining the optimal profit level, say $\tilde{\pi}$, and given this, a profit restricted slack selection problem of determining the optimal mix of slack s. Let $A(\tilde{\pi})$ be the acceptable production plans given a desired profit level of $\tilde{\pi}$, i.e.

$$A(\tilde{\pi}) = \{x \in \mathbf{R}^p_+ | \tilde{\pi} \leq b - wx\}$$

The characterization in Proposition 1 is then a special case of the characterization in Proposition 2. To see this note that the only applicable w in Proposition 2 with $A = A(\tilde{\pi})$ is the original prices w. Proposition 2 now gives us that z must solve

$$\min_{x \in L} w \cdot x$$

i.e. z must be an allocatively efficient input combination.

6 Revealed Slack Trade-Offs

The DMU's observable choice of input profile x enables us - together with an assumption of rational behavior - to make at least partial inference about the DMU's trade-off between different types of slack. We shall now demonstrate this by first deriving a series of general constraints on the DMU's value function and secondly by making explicit inference in cases with a parameterized set of possible value functions of the Weighted Fairness, Fair Gains and Nash types.

Given the technology L and the observed prices w and profits π in the price based regime, we can delineate the set of possible slack vectors available to the DMU as:

$$\begin{aligned} S &= \{s|s=(x'-z'), z' \leq x', x' \in L, z' \in L, wx' \leq b-\pi\} \\ &= (A(\pi)-L) \cap \mathbf{R}^p_+ \end{aligned}$$

In a similar way, we can use the technology L and the acceptable plans A in the restriction based regime to delineate the set of possible slack vectors that the DMU is choosing from as:

$$\begin{aligned} S &= \{s|s=(x'-z'), z' \leq x', x' \in A, z' \in L\} \\ &= (A-L) \cap \mathbf{R}^p_+ \end{aligned}$$

These choice sets are illustrated in Figure 6 below.

The actual slack choice is inferred by invoking rationality and by using our analysis above. In the price based regime we know that the rational choice of z is allocatively efficient production plan, $z \in X^a = \{\, x^a \in \mathbf{R}^p_+ | x^a \text{ solves } \min_{x' \in L} wx'\}$. The actual choice of slack must therefore be some vector

$$S^* = x - X^a.$$

In the restriction based regime, let W be the set of normed price vectors supporting x on A.[10] Moreover, let $L(W)$ be the set of efficient input combinations supported by one of the price vectors in W and dominating x, i.e. $L(W) = \bigcup_{w \in W} \{z | z \leq x, z \text{ solves } \min_{x' \in L} wx'\}$[11] .We know that a rational DMU will choose the underlying production plan z as one of the vectors from $L(W)$. The actual choice of slack must therefore be some vector

$$S^* = x - L(W)$$

The potential choice set S^* is illustrated in Figure 6 as well. Note that with strictly convex sets and unique supporting hyperplanes, these inferred choice sets are singletons.

[10] Note that a rational DMU chooses x on A by Proposition 2. If the observed x does not belong to the efficient part of A, we can say that the unit is technically inefficient, cf. the next Section.

[11] If this set is empty we can say that we have an instance of allocative inefficiency, cf. also below.

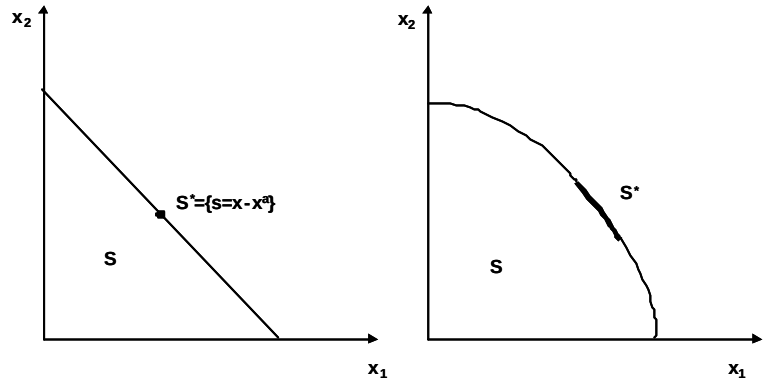

Figure 1: Figure 6: Choice of Slack

We have hereby established sets of slacks that the DMU chooses from, S, and sets of resulting potential choices, S^*. Combining this allows us to derive a series of constraints on the DMU's value function $U(\pi, s)$.

We record this as Proposition.

Proposition 3 *When the observed input use is x, the budget is b, and the prices are w or the set of acceptable plans is A, the underlying value function $U(\pi, s)$ of a rational DMU must fulfill*

$$\exists s^* \in S^* \forall s \in S : U(\pi, s^*) \geq U(\pi, s)$$

where S and S^* are defined as above.

Proof:

The proposition follows directly from the development above. □

When we make more specific assumptions about the DMU's value function and the characteristics of the technologies and acceptable input sets, we can make more precise inference. In many cases, the DMU's implicit slack aggregation can be uniquely identified. Specific cases are developed in our next Corollary. First, however, let us define some more restricted classes of slack aggregation functions.

Example 2: Instead of defining $g(\cdot)$ as an ordinary weighted average as in Example 1 and hereby implicitly assume that the slack dimensions are preferentially independent, we may use more complex weighted averages where any possible dependency between slack dimensions can be modeled through the choice of weights.

For example, we may want to assign a higher weight to slack in dimension 1 and 2 together than to slack in dimension 1 and 2 considered separately, i.e. to operate with *super-additive weights*. Hence, denote by $I = \{1, \ldots, p\}$ the slack index set and let 2^I be the set of all subsets of I. Now, define non-additive weights α on $(I, 2^I)$ as a set function $\alpha : 2^I \to [0, 1]$ such that

i) $\alpha(\emptyset) = 0, \quad \alpha(I) = 1$(normalizing the weights)

ii) Let $B, C \in 2^I$ where $B \subseteq C$ then $\alpha(B) \leq \alpha(C)$(monotonicity of weights).

By permuting the indices if necessary we may assume that slacks are increasingly ordered, i.e. $s_1 \leq s_2 \leq \ldots \leq s_p$. Given an increasing order of slacks, let $B_i = \{s_i, \ldots, s_p\}$, for any $i \in I$. With respect to non-additive weights α we may choose to aggregate slack by the Choquet-function g defined by:

$$g(s) = \sum_{i \in I} s_i(\alpha(B_i) - \alpha(B_{i+1})).$$

For further properties of this functional as well as others, for example, the Sugeno functional see e.g. Fodor and Roubens (1994).

Example 3: When inputs relate to different persons or groups , it is likely that they negotiate how to split the slack possibilities. Consequently, it may be relevant to consider a bargaining approach to the modelling of the DMU's values. An obvious candidate is the *asymmetric (Nash) bargaining value*

$$g(s) = s_1^{\alpha_1} \cdot s_2^{\alpha_2} \cdot \ldots \cdot s_p^{\alpha_p}$$

where $(\alpha_1, ..., \alpha_p) \in \mathbf{R}_+^p$, $\sum_{i=1}^p \alpha_i = 1$, represents the relative bargaining strength of the different production factors. Note that with equal strength parameters, $\alpha_i = \alpha$ for all $i = 1, \ldots, p$, we obtain the standard Nash bargaining value (see Nash(1953)). The resulting selection procedure can also be characterized by Pareto optimality, symmetry, invariance and independence of irrelevant alternatives. For further results and axiomatic characterization of the asymmetric Nash bargaining solution, see e.g. Peters (1992). □

Example 4: Another possibility is to use a slack aggregation function that focuses on the smallest weighted slack coordinate, i.e. the *weighted fairness model*

$$g(s) = \min\{\alpha_1 s_1, \alpha_2 s_2, ..., \alpha_p s_p\}$$

where $(\alpha_1, ..., \alpha_p) \in \mathbf{R}_+^p$, $\sum_{i=1}^p \alpha_i = 1$, represents the relative importance of the factors. The idea behind such a model - with all weights equal - is that the DMU is no better off than the worst of its factors. This is in line with standard Rawlsian fairness considerations, cf. Rawls(1971). Using the same notion of equality but realizing that the factors may have different alternatives, we could use a *fair gains model*

$$g(s) = \min\{s_1 - \alpha_1, s_2 - \alpha_2, .., s_p - \alpha_p\}$$

where $(\alpha_1, ..., \alpha_p) \in \mathbf{R}^p$, $\sum_{i=1}^p \alpha_i = 1$. One interpretation is that $(\alpha_1, \alpha_2, ..., \alpha_p)$, or $(\alpha_1 + t, \alpha_2 + t, ..., \alpha_p + t)$ for suitable t, represents the outside options of the different production factors. This model tries to even the gains from cooperation, i.e. the slack possibilities created in excess of what can be obtained elsewhere. □

The next Corollary shows how the Nash, Weighted Fairness and Fair Gains models can be calibrated.

Corollary 2 *In cases with i) preferential independence between slack and profit, $U(\pi, s) = V(\pi, g(s))$, ii) a unique strictly positive slack selection $S^* = \{s^*\}$ and iii) a fair weighted slack aggregation $g(s) = \min\{\alpha_1 s_1, \alpha_2 s_2, ..., \alpha_p s_p\}$, the optimal estimates of the relative importance of the factors are*

$$\alpha_i = \frac{1}{s_i(\sum_{j=1}^p s_j^{-1})} \qquad i = 1, .., p$$

In cases with i) preferential independence between slack and profit, $U(\pi, s) = V(\pi, g(s))$, ii) a unique strictly positive slack selection $S^* = \{s^*\}$ and iii) a fair allocation of gains from cooperation $g(s) = \min\{s_1 - \alpha_1, s_2 - \alpha_2, ..., s_p - \alpha_p\}$, the optimal estimates of the relative values in best alternative use are

$$\alpha_i = s_i - (1 - \sum_{i=1}^p s_i)/p \quad i = 1, .., p$$

In cases with i) preferential independence between slack and profit, , $U(\pi, s) = V(\pi, g(s))$, ii) a differentiable slack aggregation $g(.)$, and iii) a unique supporting hyperplane separating S and S^* with normal w, we have[12]

$$\frac{\partial g/\partial s_i}{\partial g/\partial s_j} = \frac{w_i}{w_j} \qquad \textit{for all } i, j = 1, .., p$$

In particular, if in the latter case, we use the Nash aggregation of slack, $U(\pi, s) = V(\pi, s_1^{\alpha_1} \cdot s_2^{\alpha_2} \cdot ... \cdot s_p^{\alpha_p})$, the bargaining power of the different factors in the Nash model can be estimated as

$$\alpha_i = \frac{w_i s_i}{\sum_{j=1}^p w_j s_j} \qquad i = 1, .., p$$

Proof:

The first part follows by noting that the most even distribution and therefore the largest value of $g(s)$ is obtained for α_i proportional to $1/s_i$. The rest is a norming.

The second part follows from noting that the most even distribution of gains is accomplished if $\alpha_i = s_i$. Now, this does not necessarily make the α values

[12] In a similar way, to estimate the trade-off between profit and aggregated slack, se could use $\frac{\partial V/\partial \pi}{\partial V/\partial g} = \frac{\partial g}{\partial s_i} \cdot \frac{1}{s_i}$ for any $i = 1, .., p$

sum to 1. We should therefore reduce (or expand) with $(t, ..., t) = te$ to hit this hyperplane. Now to get $(s + te)e = 1$ we must choose $t = (1 - \sum s_i)/p$ and we get $\alpha_i = s_i - (1 - \sum s_i)/p$.

The third part follows by the requirement that the gradient of g, $\partial g/\partial s$, must be proportional to the normal of the supporting hyperplane, i.e. to w. One way to express this is as

$$\frac{\partial g/\partial s_i}{\partial g/\partial s_j} = \frac{w_i}{w_j} \qquad i, j = 1, .., p$$

The fourth part follows form the third part, i.e. by inserting $g(s) = s_1^{\alpha_1} \cdot s_2^{\alpha_2} \cdot ... \cdot s_p^{\alpha_p}$ and using that $\partial g/\partial s_i = \alpha_i g(s)/s_i$ to obtain

$$\frac{\alpha_i g(s)/s_i}{\alpha_j g(s)/s_j} = \frac{w_i}{w_j} \qquad \text{for all } i, j = 1, .., p$$

or equivalently

$$\frac{\alpha_i}{\alpha_j} = \frac{w_i s_i}{w_j s_j} \qquad \text{for all } i, j = 1, .., p$$

Combining with $\alpha_1 + \alpha_2 + ... + \alpha_p = 1$, we get the desired result. □

The idea of the Corollary is simple. We can identify the set of possible slacks S with a given profit level or set of acceptable input vectors. If therefore, there is an optimal slack point, i.e. $S^* = \{s^*\}$ is a singleton, then the optimality of s^* over S will often contain sufficient information about the DMU's preferences to identify the unknown parameters in a class of possible value functions. The first two results illustrates this. If furthermore, not only the slack vector s^* but the slack substitution possibilities (as reflected by the price vector w) are unique, we are able to identify the slack aggregation function. The two last cases in the Corollary illustrates this.

The revelation of slack preferences through the choice of actual input combinations - and in particular the results in Corollary 2 above - are illustrated in the next example.

Example:

Consider the case with two inputs where L is the smallest convex, comprehensive set containing the input combinations $(4, 1), (2, 2)$ and $(1, 4)$. Also, assume a price-directed regime with budget $b = 8$ and prices $w = (1, 1)$. Let the actual choice of inputs be $x = (4, 3)$ corresponding to a choice of profit level equal to $\pi = b - wx = 8 - 4 - 3 = 1$. The allocatively efficient plan in this case is $x^a = (2, 2)$ such that the inferred slack consumption is $s^* = x - x^a = (2, 1)$. From the analysis in Proposition 1 we know that the set of feasible slack combinations is actually

$$S = \{s \in \mathbf{R}_+^p | s_1 + s_2 \leq 3\}$$

since the unique allocatively efficient plan is $x^a = (2, 2)$ which leaves a slack budget of $8 - 1 - 2 - 2 = 3$ when a profit of 1 is required.

The DMU's choice problem is illustrated in Figure 7 below.

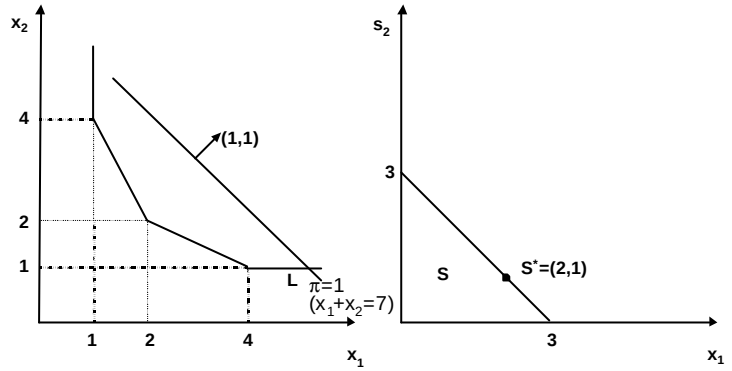

Figure 7: Choice of Slack in Example Problem

Since the DMU has chosen $s = s^* = (2, 1)$ from this set S of possible slack combinations, we can make the following claims about the underlying utility $U(\pi, s)$ for profit and slack, cf. also Proposition 3 above:

$$U(1, (2, 1)) \geq U(1, (s_1, s_2)) \qquad \forall s_1, s_2 \geq 0 : s_1 + s_2 \leq 3$$

If we furthermore assume that the underlying slack trade-off can be modeled as *weighted fairness*, i.e. by a fair weighted slack aggregation $g(s) = \min\{\alpha_1 s_1, \alpha_2 s\}$, the optimal estimates of the relative importance of the factors are

$$\alpha_1 = \frac{1}{2(\frac{1}{2} + \frac{1}{1})} = \frac{1}{3} \qquad \alpha_2 = \frac{1}{1(\frac{1}{2} + \frac{1}{1})} = \frac{2}{3}$$

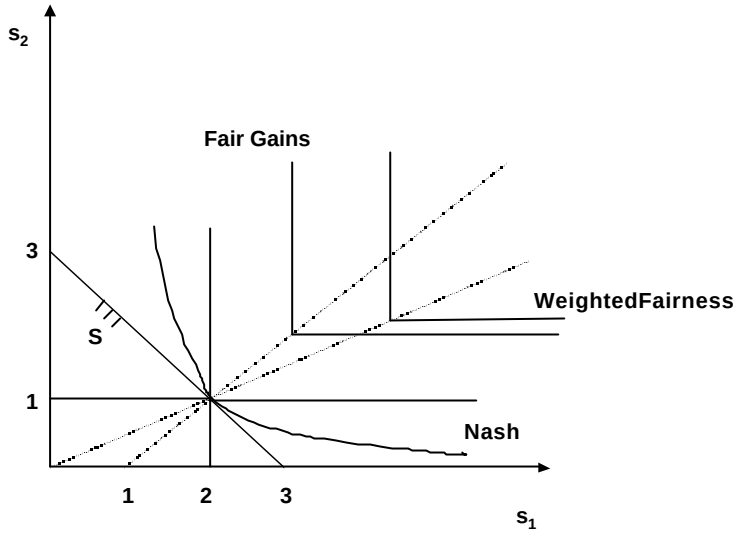

Figure 8: Revealed Slack Trade–Off in Example Problem

If we use a *fair gains model*, i.e. an equal allocation of gains from cooperation $g(s) = \min\{s_1 - \alpha_1, s_2 - \alpha_2\}$ instead, the optimal estimates of the relative values in best alternative use are

$$\alpha_1 = 2 - (1 - 2 - 1)/2 = 1 \quad \alpha_2 = 1 - (1 - 2 - 1)/2 = 0$$

Finally, if we use that the supporting prices are unique as well, we can use the *Nash bargaining model*, i.e. we can identify the Nash aggregation of slack $g(s) = s_1^{\alpha 1} \cdot s_2^{\alpha 2}$. The bargaining power of the two inputs become

$$\alpha_1 = \frac{1 \cdot 2}{1 \cdot 2 + 1 \cdot 1} = \frac{2}{3} \quad \alpha_2 = \frac{1 \cdot 1}{1 \cdot 2 + 1 \cdot 1} = \frac{1}{3}$$

These slack trade-offs are illustrated in Figure 8

7 Applications

In the previous sections, we have introduced a new perspective on inefficiency. We have suggested that inefficiency does not necessarily represent something that vanished. Rather, the excess use of resources may represent on-the-job consumption of slack and such slack may be valuable - or it may be so costly to eliminate as to make it economically inefficient to strive for technical efficiency. Moreover,

we have shown how to use observed production together with information about the technology to determine the DMU's trade-off between different types of slack. This presentation of an alternative framework and an estimation of the forces in this framework is the primary purpose of this paper.

The next natural step is to consider more specific applications of this new framework. In this section, we sketch some applications in planning, incentive provision and productivity analysis.

7.1 Planning

In the economic planning literature, cf. e.g. Bogetoft and Pruzan(1991), Dirickx and Jennergren (1979) or Johansen(1977,78), the primary means of coordinating economic units is via price plans or quantitative restrictions. The price based and restriction based regimes considered above illustrates these planning modes. If an explicit - or even a partial - model of the DMU's value structure $U(\pi, s)$ can be estimated, one can make rather precise predictions of the rational response to such controls.

To formalize this, let us assume that the planner uses price control to set the factor prices, e.g. transfer prices, to $\overline{w}$ and that he uses restrictions to limit the use of inputs to $\overline{X}$ and the budget to $\overline{b}$. Using an estimated value model $U(\pi, s)$, we can now predict the response by the DMU. The rational response of the DMU is the solution to the value maximization problem

$$\begin{array}{ll} \max & U(\overline{b} - \overline{w}x, x - z) \\ z, x & \\ \text{st} & \overline{w}x \leq \overline{b} \\ & z \leq \underline{x} \\ & x \in \overline{X}, z \in L, \end{array}$$

To illustrate the use of an inferred value of slack model, consider the example from the last section. If we consider variations in the prices, we see from Figures 7 and 8 that a general shift in budget or in the price level (but not in the relative prices) will lead to a linear expansion or contraction in the Nash and Weighted Fairness Models. Movements in the Fair Gains Model will however be along the slack line through (1,0) and (2,1). In terms of input vectors, this means that expansions / contractions in the first cases will happen along the line through (2,2) and (4,3) while it will happen along the line through (2,2), (3,2) and (4,3) in the latter case. Figure 9 illustrates this.

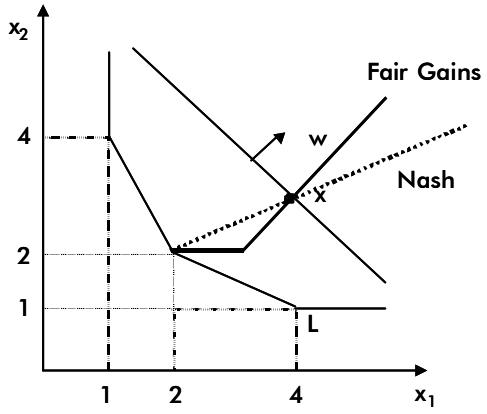

Figure 9: Reaction Paths to Varying Budget Restrictions

Note also that if we change the relative prices but not the budget needed for the original slack combination (2,1), we see that for variations that do not alter the allocative efficiency of $z = (2, 2)$, the response in the Nash Model will be to favor the less costly input while there will be no substitution in the Weighted Fairness and Fair Gains Models. In similar ways, one can analyze the rational responses to new types of restrictions on input usage.

7.2 Incentives

When prices are changed or restrictions are introduced on the budget or - more generally - on the input usage directly, it is important to take into account the incentive effects on the DMU in terms of both direct off-the-job profit and indirect on-the-job slack. More specifically, it is important to respect the *individual rationality* of the DMU. The DMU, whether a single or a multiple agents organization, may require a certain minimal value to continue. Similarly, it is important to respect the *incentive compatibility* constraints. If a particular behavior, i.e. the choice of a particular underlying production plan z and a particular slack plan s, is intended then it must be in the best interest of the DMU itself to pick this within the planning regime.

To illustrate incentive provision, consider the planner's problem of designing the price and budget levels to minimize the cost of inducing the DMU to produce the desired output y. If the actual cost to the planner of providing inputs x is

$c(x)$, his contract design problem reads:

$$\begin{array}{lll} \min & b + c(x) - wx & \\ b, x, w, z & & \\ \text{st} & U(b - wx, x - z) \geq \overline{U} & IR \\ & z \in \arg\min_{z' \in L} wz' & IC - z \\ & U(b - wx, x - z) \geq U(b - wx', x' - z) \quad \forall x' \geq z & IC - x \end{array}$$

The objective in this program is to minimize the cost of resources $c(x)$ plus the net-payment $b - wx$ to the DMU. The first constraint is the individual rationality constraint IR. It requires that the value to the DMU from selecting the intended plan is at least its reservation utility. The next constraint is part of the needed incentive compatibility constraints. $IC - z$ uses Proposition 1, i.e. the optimality of choosing the underlying production plan as an allocatively efficient one. The third set of incentive compatibility constraints, $IC - x$, makes it a best response for the DMU to make a total use x of the inputs.

In practice, we may not know the reservation value of a DMU. We may however assume that it was fulfilled in the last period and try to make reallocations that does not reduce the resulting value to the DMU. To illustrate this, assume that the budget, prices and inputs used in the last period was $\tilde{b}, \tilde{w}$, and $\tilde{x}$. The planner can therefore in the program above use $\overline{U} := U(\tilde{b} - \tilde{w}\tilde{x}, \tilde{x} - z(\tilde{w}))$, where $z(\tilde{w})$ is an allocatively efficient input vector with prices $\tilde{w}$.

Even more fundamentally, we may assume that the planner does not know the trade-offs between slacks nor the trade-offs between slack and profit. That is, he only knows that there is some underlying function $U(\pi, s)$ that increases in profit and slack. Although he now has very limited information, his possibilities to suggest welfare improving reallocations are not entirely crippled. He may simply ensure that neither the profit level nor the set of slack possibilities are reduced. To determine the optimal such profit and slack preserving reallocation, he could solve

$$\begin{array}{lll} \min & b - wx & \\ b, x & & \\ \text{st} & b - wx \geq \tilde{b} - \tilde{w}\tilde{x} & IR - \pi \\ & S(x) \supseteq S(\tilde{x}) & IR - s \\ & x \in L & \end{array}$$

Here, we have assumed again that the planning can only take place using a budget and a demand for output y. The rest is left for the DMU. More advanced planning instruments (transfer prices deviating from market prices to the planner, quantitative restrictions etc.) could be looked for like above.

7.3 Productivity Measures

Although we have developed a framework that can rationalize several instances of inefficiency, there are still behavioral patterns that represent "waste". In other

words, it is possible to test and reject the rationality hypothesis. To see this, consider the restriction based regime depicted in Figure 10 below.

Three production plans, x^1, x^2 and x^3, are shown, all of which involve irrational waste of resources.

In the first production plan, x^1, the DMU does not consume all the available slack given the acceptable set A. It could move north-east and get more of all slacks. We could call this *organizational inefficiency* in the new framework.[13]

In the second production plan, x^2, the choice of inputs is not compatible with the hypothesis of maximizing slacks given input prices which result in the cost minimizing allocation x^a. Whatever underlying production plan $z \leq x^2$ is associated with x^2, the DMU is left with less of both slacks than what can be accomplished by using the allocatively efficient plan $z = x^a$ and an appropriate x on the "efficient" part of A above z, i.e. x such that $x \geq x^a$ and $x \in D(A)$. Production in x^2 can be called *allocatively inefficient* since the DMU has not managed to allocate the inputs so as to maximize the slack potentials.

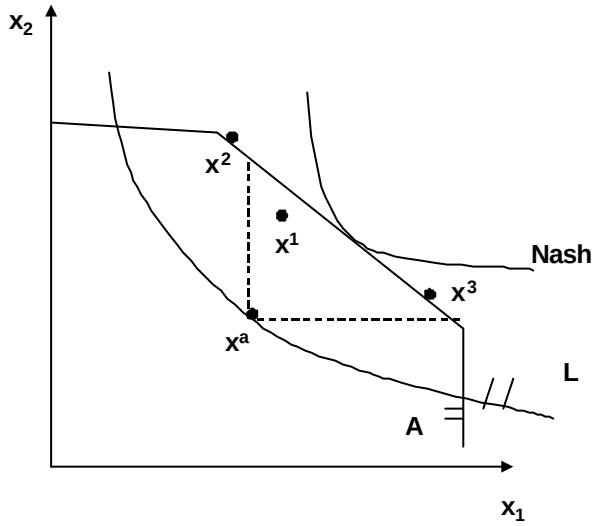

Figure 10: Irrational Inefficiencies

The last production plan, x^3, may be allocatively efficient, namely if the slack trade-off favors a large slack in the first dimension. If however, the slack aggregation is known to follow a more specific model, say a Nash model with equal

[13]From the point of view of the DMU, we could even call it technical inefficiency since it implies that the DMU has not moved to the frontier of its slack possibility set.

bargaining strengths, x^3 can be termed inefficient as well. In this case, we have an instance of preferentially allocative inefficiency - or simply *value inefficiency* - in the sense that the DMU has not managed to take full account of its - in another context - revealed slack trade-offs.

In all three cases, it is straightforward to use an estimated value function to measure the "value" of the inefficiency. The difficult part is to estimate the value function properly.

There are of course many other ways to use the present framework in productivity analysis. To illustrate, the observation of x^2 could be rationalized if the technology was different. We can therefore stick to the presumed rationality and instead use the observation of x^2 to extend the technology L. Another development would be to introduce non-parametric models of the values of slack and to estimate these - together with the technology - using the observed productions. It may also be worthwhile to introduce a notion of "the least costly efficiency improvement" by determining the path from a given x to $F(L)$ that marginally reduces the value of the slack the least.[14]The contraction path developed for our numerical example in Section 7.1 illustrates this.

8 Final Remarks

The aim of this paper has been to introduce a view of inefficiency as potentially rational. We have modeled this by introducing an aggregate description of a DMU's intention as one of maximizing both profits and slack. From observed production plans, we can make at least partial inference about the DMU's trade-off between on-the-job consumption of slack and off-the-job consumption of profit as well as between the consumption of different types of slack. We have also sketched some possible applications of this perspective within planning, incentive provision and productivity measurements.

From the point of view of productivity analysis, the crucial feature of our approach is that we look at the DMU from the "inside". In this respect, we challenge the traditional black-box view of inefficiency underlying Leibenstein(1966)'s X-efficiency which is, in fact, used in much of the modern literature on productivity analysis. We follow the suggestions of Stigler(1976) and see the task as one of explaining why differences in efficiency may occur. Once understood and modelled we are more qualified to deal with issues of efficiency. From the DMU's perspective, there are many reasons to have inefficiency. We try to capture several of these reasons in a model that can be at least partially calibrated on observed

[14]It may be useful to compare our perspective to the literature combining Data Envelopment Analysis and Multiple Criteria Decision Making. In particular the notion of value efficiency introduced by Halme e.a.(1999) is interesting here and the two approaches do share some similarities although our approach relies only on implicit value revelation based on observed productions and the value efficiency approach relies on additional, explicit value elicitation.

data. Hereby, the approach becomes operational and may be used to improve planning, incentive provision and productivity measurements. In other words, we do not solely try to explain why differences may occur.. In fact, we use rather stylized organizational perspectives. We seek also to measure the forces involved, including the relative value of different types of slack in the DMU, and to suggest how this can affect the "outside" view of organizations used in traditional productivity analysis.

There are many ways to extend the approach suggested here. From the point of view of developing our framework further, however, we suggest that in particular three avenues will be productive. First it is worthwhile to extend and refine the still stylized view of organizational behavior adopted here. It is worthwhile for example to expand on the different uses of slack and to consider alternative control regimes. Secondly, it is important to develop more theoretical "applications" in greater details. It is worthwhile for example to re-think basic concepts like allocative efficiency in this new framework and to consider how to make specific efficiency measurements using information about the trade-off between slacks. Also, it relevant and most likely possible to combine with models of satisficing behavior, with other (and most challenging some non-parametric) bargaining models, with uncertainty models and with some alternative regulatory models. Thirdly, it would be interesting to estimate some of the more specific models on actual data sets and see how it would modify more specific planning, incentive or productivity measurement problems. We leave all of this for future research.

References

[1] Afriat, S., Efficiency Estimation of Production Functions, *International Economic Review*, 13, pp.568-598, 1972..

[2] Banker, R.D. and A Maindiratta, Nonparametric Analysis of Technical and Allocative Efficiencies in Production, *Econometrica,* 56, pp.1315-1332, 1988.

[3] Bogetoft, P., Non-Cooperative Planning Theory, pp.1-314, Springer-Verlag, 1994a.

[4] Bogetoft, P., Incentive Efficient Production Frontiers: An Agency Perspective on DEA, *Management Science*, 40, pp.959-968, 1994b.

[5] Bogetoft, P., Incentives and Productivity Measurements, *International Journal of Production Economics*, 39, pp. 67-81, 1995.

[6] Bogetoft, P., DEA-Based Yardstick Competition: The Optimality of Best Practice Regulation, *Annals of Operations Research*, 73, pp. 277-298, 1997.

[7] Bogetoft, P., DEA and Activity Planning under Asymmetric Information, *Journal of Productivity Analysis*,13, pp. 7-48, 2000.

[8] Bogetoft, P. and P.Pruzan, Planning with Multiple Criteria: Investigation, Communication and Choice, North-Holland, 1991.

[9] Chavas, J.P. and T.L.Cox, A Nonparametric Analysis of Productivity: the Case of U.S. and Japanese Manufacturing, *American Economic Review*, 80, pp.450-464, 1990.

[10] Chavas, J.P and T.L.Cox, On Generalized Revealed Preference Analysis, *Quarterly Journal of Economics,* 74, pp. 593-571, 1992.

[11] Chavas, J.P and T.L.Cox, On Non-Parametric Supply Response Analysis, *American Journal of Agricultural Economics,* 77, pp.80-92, 1995.

[12] Cyert, R.M. and J.G. March, A Behavioral Theory of the Firm, Prentice Hall, 1963.

[13] Debreu, G., The Coefficient of Resource Utilization, *Econometrica*, 19, pp. 273-292, 1951.

[14] DeAllessi, L., An Economic Analysis of Government Ownership and Regulation: Theory and the Evidence from the Electric Power Industry, *Public Choice*, 19, pp. 1-42, 1974.

[15] Dirickx, Y.M.I. and L.P. Jennergren, Systems Analysis by Multilevel Methods, John Wiley & Sons, 1979.

[16] Farrell, M.J., The measurement of productive efficiency, *Journal of the Royal Statistical Society*, Series A, III, pp. 253-290, 1957.

[17] Färe, R, Fundamenyals of Production Theory, Springer-Verlag, 1988.

[18] Färe, R. and S.Grosskopf, Non-Parametric Tests of Regularity, Farrell Efficiency, and Goodness of Fit, *Journal of Econometrics,* 69, pp.415-425, 1995.

[19] Fodor, J. and M. Roubens, Fuzzy Preference Modelling and Multicriteria Decision Support, Klüwer Academic Publishers, 1994.

[20] Galbraith, J.R,, Organizational Design: An Information Processing View, *Interfaces*, 4, pp.28-36, 1974.

[21] Grosskopf, S. and K.Hayes, Local Public Sector Bureaucrats and Their Input Choices, *Journal of Urban Economics*, 33, pp.151-166, 1993.

[22] Halme, M., T.Joro, P.Korhonen, S.Salo and J.Wallenius, A Value Efficiency Approach to Incorporating Preference Information in Data Envelopment Analysis, *Management Science*, 45, pp. 103-115,1999.

[23] Hanoch, G. and M. Rothschild, Testing Assumptions of Production Theory: A Nonparametric Approach, *Journal of Political Economy,* 80, pp.256-275, 1972.

[24] Haskel, J. and A.Sanchis, Privatisation and X-inefficiency: A Bargaining Approach, *Journal of Industrial Economics*, 43, pp.301-321, 1995.

[25] Haskel, J. and A.Sanchis, A Bargaining Model of Farrell Inefficiency, *International Journal of Industrial Organization*, 18, pp.539-556, 2000.

[26] Johansen, L., Lectures on Macroeconomic Planning, Part I: General Aspects, North-Holland, 1977.

[27] Johansen, L., Lectures on Macroeconomic Planning, Part II: Centralization, Decentralization, Planning under Uncertainty, North-Holland, 1978.

[28] Koopmans, T.C., Analysis of Production as an Efficient Combination of Activities, in T.C.Koopmans(ed.), Activity Analysis of Production and Allocation, New York, Wiley, 1951.

[29] Lindsay, C., A Theory of Government Enterprise, *Journal of Political Economy*, 84, pp.1061-1077, 1976.

[30] Leibenstien, H., Allocative efficiency vs. 'X-efficiency', *The American Economic Review*, 56, 392-415, 1966.

[31] Leibenstein, H., X-inefficiency Xists: Reply to an Xorcist, *The American Economic Review*, 68, 203-211, 1978.

[32] Leibenstein, H. and S. Maital, The Organizational Foundation of X-inefficiency: A Game Theoretic Interpretation of Argyris' Model of Organizational Learning, *Journal of Economic Behavior and Organization*, 23, 251-268, 1994.

[33] Migué, J.L. and G.Bélanger, Towards a General Theory of Managerial Discretion, Public Choice, 17, pp.27-43, 1974.

[34] Niskanen, W.A., Bureaucracy and Representative Government, Aldine Press, 1971.

[35] Parish, R. and Y-K Ng, Monopoly, X-efficiency and the Measurement of Welfare Loss, *Economica*, 39, pp.301-308, 1972.

[36] Peters, H.J.M, Axiomatic Bargaining Game Theory, Klüwer Academic Publishers, 1992.

[37] Rawls, J., A Theory of Justice, Harvard University Press, 1971.

[38] Samuelson, P., Foundations of Economic Analysis, Havard University Press, Cambridge, Massachusetts, 1947.

[39] Stabler, U. and J. Sydow, Organizational Adaptive Capacity: A Structuration Perspective, *Journal of Management Enquiry*, to appear, 2001.

[40] Stigler, G. , The Xistence of X-efficiency, *The American Economic Review*, 66, 213-216, 1976.

[41] Shephard, R.W., Cost and Production Functions, Princeton University Press, 1970.

[42] Tullock, G. The Welfare Costs of Tariffs, Monopolies and Theft, *Western Economic Journal*, 5, 224-232, 1967.

[43] Varian, H.R., The Non-Parametric Approach to Production Analysis, *Econometrica*, 52, pp. 279-297, 1984.

[44] Varian, H.R., Non-Parametric Tests of Optimizing Behavior with Measurement Error, *Journal of Econometrics,* 30, pp.445-458, 1985.

[45] Varian, H.R., Revealed Preference with a Subset of Goods, *Journal of Economic Theory*, 46, pp.179-185, 1988

[46] Varian, H.R., Goodness-of-Fit in Optimizing Models, *Journal of Econometrics*, 46, pp.125-140, 1990.

[47] Weiss, A., Efficiency Wages: Models of Unemployment, Layoffs and Wage Dispersion, Princeton University Press, 1990.

[48] Williamson, O., The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm, Printice

[49] -Hall, 1964.